ProPILE: Probing Privacy Leakage in Large Language Models

Year: 2023 | Citations: 159 | Authors: Siwon Kim, Sangdoo Yun, Hwaran Lee

---

Abstract

The rapid advancement and widespread use of large language models (LLMs) have raised significant concerns regarding the potential leakage of personally identifiable information (PII). These models are often trained on vast quantities of web-collected data, which may inadvertently include sensitive personal data. This paper presents ProPILE, a novel probing tool designed to empower data subjects, or the owners of the PII, with awareness of potential PII leakage in LLM-based services. ProPILE lets data subjects formulate prompts based on their own PII to evaluate the level of privacy intrusion in LLMs. We demonstrate its application on the OPT-1.3B model trained on the publicly available Pile dataset. We show how hypothetical data subjects may assess the likelihood of their PII being included in the Pile dataset being revealed. ProPILE can also be leveraged by LLM service providers to effectively evaluate their own levels of PII leakage with more powerful prompts specifically tuned for their in-house models. This tool represents a pioneering step towards empowering the data subjects for their awareness and control over their own data on the web.